BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 6 2010, pages 752–760
doi:10.1093/bioinformatics/btq043

Structural bioinformatics

Advance Access publication February 3, 2010

Cascleave: towards more accurate prediction of caspase substrate cleavage sites

Jiangning Song[1,2,*,†], Hao Tan[1,3,†], Hongbin Shen[4], Khalid Mahmood[1,5], Sarah E. Boyd[3], Geoffrey I. Webb[3], Tatsuya Akutsu[2,*] and James C. Whisstock[1,5,*]

[1]Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia, [2]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan, [3]Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia, [4]Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200240, China and [5]ARC Centre of Excellence for Structural and Functional Microbial Genomics, Monash University, Melbourne, VIC 3800, Australia

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The caspase family of cysteine proteases play essential roles in key biological processes such as programmed cell death, differentiation, proliferation, necrosis and inflammation. The complete repertoire of caspase substrates remains to be fully characterized. Accordingly, systematic computational screening studies of caspase substrate cleavage sites may provide insight into the substrate specificity of caspases and further facilitating the discovery of putative novel substrates.

Results: In this article we develop an approach (termed Cascleave) to predict both classical (i.e. following a P_1 Asp) and non-typical caspase cleavage sites. When using local sequence-derived profiles, Cascleave successfully predicted 82.2% of the known substrate cleavage sites, with a Matthews correlation coefficient (MCC) of 0.667. We found that prediction performance could be further improved by incorporating information such as predicted solvent accessibility and whether a cleavage sequence lies in a region that is most likely natively unstructured. Novel bi-profile Bayesian signatures were found to significantly improve the prediction performance and yielded the best performance with an overall accuracy of 87.6% and a MCC of 0.747, which is higher accuracy than published methods that essentially rely on amino acid sequence alone. It is anticipated that Cascleave will be a powerful tool for predicting novel substrate cleavage sites of caspases and shedding new insights on the unknown caspase-substrate interactivity relationship.

Availability: http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/Cascleave/

Contact: jiangning.song@med.monash.edu.au; takutsu@kuicr.kyoto-u.ac.jp; james.whisstock@med.monash.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 24, 2009; revised and accepted on January 28, 2010

1 INTRODUCTION

To date 14 mammalian caspases have been identified; these molecules function in signaling cascades that control critical processes such as apoptosis, necrosis, inflammation, migration proliferation and differentiation (Dix *et al*., 2008; Fischer *et al*., 2003; Lüthi and Martin, 2007; Mahrus *et al*., 2008; Nicholson, 1999; Pop and Salvesen, 2009; Talanian *et al*., 1997; Timmer and Salvesen, 2007). Caspases cleave substrates that generally contain a highly conserved aspartate (D) at the P_1 position (Nicholson, 1999; Pop and Salvesen, 2009; Talanian *et al*., 1997). However, in addition to the P_1 Asp, the amino acids at the P'_1 and P_4–P'_2 positions contain important additional determinants of specificity that can dramatically affect cleavage efficiency. Accordingly, mammalian caspases can be divided into three groups (Nicholson, 1999). Group I caspases (caspase-1, -4, -5 and -13) prefer bulky hydrophobic amino acids at the P_4 site and recognize peptide sequence (W/L)EHD; group II caspases (caspase-2, -3 and -7) preferentially cleave the DEXD motif, while group III caspases (caspase-6, -8, -9 and -10) recognize the sequence (I/V/L)E(H/T)D. In contrast to the other caspases, caspase-14 is expressed and activated mainly in the epidermis and exhibits cleavage preference for the WEHD or IETD motif (Denecker *et al*., 2008).

The study of apoptotic pathways predominantly mediated by caspases has important implications for the development of therapies for cancer treatment, there is significant interest in gaining a better understanding of caspase substrate specificity (Dix *et al*., 2008; Lüthi and Martin, 2007; Mahrus *et al*., 2008; Timmer *et al*., 2009). Even though almost 400 caspase substrates have been reported to date, there are likely to be hundreds of new caspase substrates that remain to be discovered (Fischer *et al*., 2003). Experimental identification and characterization of protease substrates is often difficult and time-consuming (Enoksson and Salvesen, 2008; Enoksson *et al*., 2007; Ju *et al*., 2007; Rawlings *et al*., 2008; Schilling and Overall, 2008). Hence, computational prediction of caspase substrate specificity may provide useful and experimentally testable information in regards to novel potential cleavage sites or candidate substrates.

Several computational approaches have been developed to predict caspase cleavage-site specificity. PeptideCutter utilized a limited experimental dataset and was initially used to predict the substrate cleavage sites for a variety of protease families including several

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

caspases (Gasteiger *et al*., 2005). Lohmüller *et al*. (2003) developed a bioinformatic tool called prediction of endopeptidase substrates (PEPS) that utilized rule-based endopeptidase cleavage site scoring matrices (CSSM) and was deployed to predict caspase-3 substrates. Garay-Malpartida *et al*. built CasPredictor to predict caspase cleavage sites; this approach again utilizes sequence and specificity information but also incorporates an index of PEST-like sequences enriched in the vicinity of cleavage site regions (Garay-Malpartida *et al*., 2005; Singh, 2006). The latter program achieved an accuracy of 81% when evaluated on a dataset of 137 experimentally verified cleavage sites (Garay-Malpartida *et al*., 2005). Yang (2005) applied neural networks to build models for predicting caspase cleavage sites and investigated the impact of a sliding window size on the performance. This approach attained the highest reported prediction accuracy of 97%, however, it is important to note that only a small dataset of 13 substrate sequences was considered. The GraBCas software developed by Backes *et al*. (2005) provided position-specific scoring prediction of cleavge sites for caspases 1–9 and granzyme B. Wee *et al*. applied a support vector machine (SVM) approach and showed that CASVM predictors could achieve an accuracy ranging from 81.2 to 97.9% based on an extended dataset of 210 cleavage sites. More recently, they proposed a two-step model called multi-factor CASVM which exploits the structural factors to score and filter the false positives to improve the performance (Wee *et al*., 2009).

In this work, we extend the SVM approach proposed by Wee *et al*. (2006, 2007, 2009) to predict caspase cleavage sites from the flanking sequences of substrates. We started with the construction of a caspase substrate database with experimentally verified substrate sequences extracted from multiple resources. We then built support vector regression (SVR) models that not only provided conventional two-state (cleavage or non-cleavage) prediction, but also generated an estimated probability for each candidate cleavage site, thus providing a quantitative evaluation of caspase substrate specificity. We extensively explored different sequence encoding schemes and examined their effects on the prediction performance. Further, we took into consideration the characteristic sequence and structural features surrounding substrate cleavage and non-cleavage sites, such as the predicted secondary structure, solvent accessibility and natively disordered regions, based on a recently developed bi-profile Bayesian feature extraction method (Shao *et al*., 2009). Comparison with the published approaches reveals that our approach is generally useful for identification of caspase cleavage sites in large datasets.

2 METHODS

2.1 Datasets

We have constructed a caspase substrate sequence dataset from multiple resources, including manually curated CASBAH database (Fischer *et al*., 2003), the MEROPS database (Rawlings *et al*., 2008), the CASVM webserver (Wee *et al*., 2006, 2007), the Uniprot database (Bairoch and Apweiler, 2000), as well as a literature search. All the annotated substrate cleavage sites were verified experimentally. The current dataset contains 370 caspase substrate sequences and 562 cleavage sites (Supplementary Table 1). In order to objectively evaluate the prediction performance, we employed 5-fold cross-validation and leave-one-out cross-validation (LOOCV) methods. In the case of 5-fold cross-validation, substrate sequences in this dataset were randomly divided into five subsets with roughly equal numbers of substrate sequences. In each validation step, one subset was singled out in turn as the testing dataset, while the rest were used as the training dataset. In the case of LOOCV, each substrate sequence in the dataset was singled out in turn as the testing set and this procedure was repeated for all the substrate sequences. Moreover, in order to make a more stringent comparison with CASVM, we also used an independent testing set extracted from a recent experimental study (Dix *et al*., 2008), which contains 64 caspase substrate sequences and 69 cleavage sites. The complete list of these substrate sequences used in this study is available in the Supplementary Data and can be downloaded from the Cascleave website (http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/Cascleave/).

Cleavage and non-cleavage sites were collected as positive and negative datasets, respectively. Peptide sequences in positive and negative datasets were extracted using a sliding window approach surrounding the experimentally verified cleavage sites and other residues that were found not to be cleaved by caspases, respectively. The predictive models were then built based on the extracted positive and negative peptide sequences. As the obtained datasets this way might contain sequence redundancy which will lead to the overestimation of the prediction performance of the models, we need to reduce sequence homology for the extracted positive and negative peptide sequences. Local window-based homology reduction (Shao *et al*., 2009) was adopted for this purpose. Sequence homology reduction within the training and testing datasets was performed in such a way that sequence identity between any two peptide sequences should not be larger than 70%.

2.2 Binary encoding amino acid sequence profiles (BEAA)

In the first instance, the SVR models were trained and tested using binary encoding amino acid sequence profiles (BEAA) where substrate sequences were transformed into *n*-dimensional vectors using an orthonormal encoding scheme, in which each amino acid is represented by the 20-D binary vector composed of either zero or one elements (Song *et al*., 2006; Wee *et al*., 2006), e.g. Ala (10000000000000000000), Cys (01000000000000000000), Asp (00100000000000000000), …, Tyr (00000000000000000001), etc. For the sake of simplicity, we termed this binary encoding amino acid sequence profile as the encoding scheme 'BEAA'. Since increasing sequence window size is supposed to provide more local sequence information, we used a sliding window approach to derive the local sequence profiles based on the 'BEAA' scheme and examined the corresponding prediction accuracy. The window size w is defined as the residue numbers involved in the local sequence windows surrounding the cleavage sites from P_8 to P'_8 positions, either in a symmetrical or non-symmetrical manner (Supplementary Fig. 1): i.e. $w = 3$ (P_4–P_1), 4 (P_2–P'_2 or P_3–P'_1), 5 (P_3–P'_2 or P_4–P'_1), 6 (P_3–P'_3), …, 14 (P_7–P'_7), 16 (P_8–P'_8), etc.

2.3 Predicted structural information

Based on the observation of structural determinants of caspase substrate specificity, we also incorporated into Cascleave the structural information predicted by state-of-art algorithms, specifically, secondary structures, solvent accessibility and natively unstructured regions. Secondary structure was predicted using the PSIPRED program, which provides one of the most accurate predictions for protein secondary structures and generates the probability profiles of three secondary structure assignments (helix, strand and coil) for each residue in a protein (Jones, 1999). For a given residue, we extracted the $w \times 3$ matrix from the output file of PSIPRED by selecting the sliding window size w. Solvent accessibility was predicted using the SSpro program implemented in the SCRATCH package (Cheng *et al*., 2005). SSpro predicts the solvent accessibility status for each residue in a protein sequence, producing a binary output- either as 'exposed' or 'buried' (Cheng *et al*., 2005). Solvent accessibility was encoded as binary units into the SVR model. Natively unstructured region was predicted using the DISOPRED2 server (Ward *et al*., 2004), which is one of the leading servers for predicting natively disordered regions in proteins. The probability of each residue being disordered generated by DISOPRED2 is used as the input to the SVR models.

2.4 Bi-profile Bayesian signatures

Shao *et al.* (2009) recently proposed a novel approach called bi-profile Bayes and applied it to predict methylation sites in proteins. This approach has been demonstrated to provide a significant improvement of prediction performance for predicting methylation sites from protein sequences (Shao *et al*., 2009). We will describe the application of bi-profile Bayesian feature extraction approach to predict caspase cleavage sites in this study. The rationale behind this approach is that peptide sequences that can be cleaved by caspases should exhibit different features or characteristics relative to those that cannot be cleaved. Therefore, integrating the bi-profile Bayesian signatures by representing each sample in a bi-feature manner would be more informative than the single BEAA mentioned above. This approach would be particularly useful when dealing with an unbalanced dataset comprising a smaller amount of positive samples and greater number of negative samples. Given a substrate peptide sequence $S = \{aa_1, aa_2, aa_3, \ldots, aa_i, \ldots, aa_n\}$, where each aa_i ($i = 1, \ldots, n$) denotes an amino acid at position i and n denotes the sequence length of substrate peptide (i.e. the local sliding window size), S can be classified as one of the two classes C_1 (representing cleavage sites of caspases, i.e. positive samples) or C_0 (representing non-cleavage sites of caspases, i.e. negative samples), according to whether or not it can be cleaved by caspases. The posterior probability of both positive and negative samples can be calculated as the occurrence of each amino acid at each position in the training dataset. We integrated the bi-profile Bayesian signatures to predict cleavage sites of caspase from primary sequence of substrates (for more details about the bi-profile Bayesian feature extraction method, refer to the original work of Shao *et al*.).

2.5 Sequence encoding schemes

We extracted four different types of sequence-profiles based on the bi-profile Bayesian feature extraction approach: (i) bi-profile Bayesian amino acid profile (BPBAA); (ii) bi-profile Bayesian secondary-structure profile (BPBSS); (iii) bi-profile Bayesian solvent accessibility profile (BPBSA); (iv) bi-profile Bayesian disordered profile (BPBDISO), through calculating the frequency of each amino acid or the corresponding structural types at each position for the cleavage and non-cleavage peptide sequences in the training set. As we performed 5-fold/LOOCV, in each of cross-validation rounds, these bi-profile Bayesian signatures were generated based only on the training dataset without the validation subset used as the testing dataset, in order to avoid the overestimation of the prediction performance. Based on the bi-profile Bayesian signature extraction, a peptide sequence (either positive or negative) can be encoded by its probability vector $\vec{P} = (p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_{2n-1}, p_{2n})$, potentially carrying more informative features than the simple binary encoding amino acid (BEA) profiles.

2.6 SVR implementation and parameter selection

In this study, we used SVR to build the models to estimate the cleavage probability of caspase substrates. SVM is a supervised machine learning technique based on structural risk minimization from statistical learning theory (Vapnik, 2000). SVM has been applied successfully to a wide range of classification problems in recent years, including predicting protein subcellular localization (Kumar and Raghava, 2009; Tamura and Akutsu, 2007), protein–protein interaction (Shen *et al*., 2007), protein methylation sites (Shao *et al*., 2009), DNA-repair protein (Brown and Akutsu, 2009), cyclin protein (Kalita *et al*., 2008) and microRNAs (Ahmed *et al*., 2009). In practice, SVM has two modes: the classification mode SVC and regression mode SVR. Due to its excellent regression ability to infer property values from a limited dataset of samples, SVR has attracted increasing attention with a growing number of applications, including predicting gene expression level (Raghava and Han, 2005), residue contact number (Yuan, 2005), residue contact order (Song and Burrage, 2006), residue depth (Song *et al*., 2009), peptide-MHC binding affinities (Liu *et al*., 2006; Wan *et al*., 2006), disulfide connectivity (Song *et al*., 2007) and half-sphere exposure (Song *et al*., 2008). We used the SVM_light package, an implementation of Vapnik's SVM for SVC, SVR and pattern recognition (Joachims, 1999). We selected radial basis kernel function (RBF) at $\varepsilon = 0.01$, $\gamma = 0.01$ and $C = 100.0$ to build the prediction models. This parameter set was optimized based on 5-fold cross-validation.

The number of negative samples is much larger than that of the positive samples, which will incur the imbalance problem and result in biased prediction in favor of the negative data (Song *et al*., 2006; Wee *et al*., 2006). We employed the under-sampling approach to overcome this imbalance problem by reducing the size of the over-represented negative samples. We set the ratio of the positive to negative data at 1:3. Random sampling still retained the original distribution of negative samples, but avoided their over-representation.

2.7 Performance evaluation

The predictive performance of our approach was evaluated using the accuracy, sensitivity, specificity, F-score and Matthews correlation coefficient (MCC) measures (Matthews, 1975) (see Supplementary Data for more details). The predictive performances were evaluated using these five measures based on 5-fold cross-validation and LOOCV tests.

3 RESULTS

3.1 Statistical distribution of substrate cleavage sites

Using the compiled dataset of experimentally determined caspase substrates, we analyzed the statistical distributions of substrate cleavage sites for P_8–P'_8 positions. As shown in Figures 1 and 2, caspase-3 was used as an example to generate a heat map and sequence logo diagram, respectively, for the P_8–P'_8 specificities of caspase-3. As expected, one of the prominent characteristics of caspase-3 cleavage site specificity is that this enzyme preferentially cleaves after the aspartate (D) at P_1 and P_4 position, which forms the well-known canonical 'XXXD' motif (Dix *et al*., 2008; Fischer *et al*., 2003; Lüthi and Martin, 2007; Mahrus *et al*., 2008; Nicholson, 1999; Pop and Salvesen, 2009; Talanian *et al*., 1997; Timmer and Salvesen, 2007; Timmer *et al*., 2009).

Our analysis reveals that 96.99% of cleavage sites have a P_1 aspartate, while only a small percentage of cleavages sites exhibit the non-canonical 'XXXE' (1.24%), 'XXXG' (0.53%) and 'XXXA' (0.35%) motifs. Aside from the P_1 site specificity, we note a modest preference for glutamate at P_3 and glycine at P'_1. Aspartate was present in 43.44% of all P_4 positions and glycine occurred in 28.19%

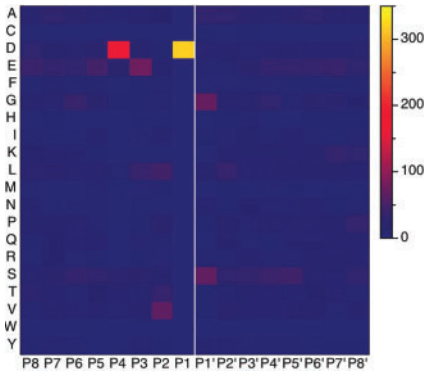

Fig. 1. Caspase-3 substrate cleavage sites for P_8–P'_8 position. The average amino acid occurrences in P_8–P'_8 were calculated and displayed in the form of a two-dimensional heat map. The scissile peptide bond between sites P1 and P1′ is indicated by a vertical white line.

Fig. 2. Sequence logo diagram representation of the occurrences of amino acid residues in the caspase substrate cleavage site P_8–P'_8 positions. To better show the occurrence rate of each amino acid type, the sequence logo ordinates have been scaled in bits (Schneider and Stephens, 1990).

of all P'_1 positions. We also generated heat maps for the other caspases such as caspase-1, -6, -7, -8, -9 and -10, as well as sequence logo representations for P_8–P'_8 position (Supplementary Figs 1–3). Comparison of different caspase groups reveals distinct patterns of subsite specificity through the P_8–P'_8 sites, for example, serine at P'_1 for caspase-1, valine at P_4 for caspase-6, glycine at P'_1 for caspase-7 and proline at P_8 for caspase-10.

3.2 Analysis of structural determinants that characterize caspase substrate specificity

As indicated in previous studies, the three-dimensional (3D) context and the appropriate presentation of solvent accessible surface are key factors, which determine whether the presence of a particular substrate motif can be accessed and cleaved by a caspase (Nicholson, 1999; Timmer *et al*., 2009). The expanded dataset of caspase substrates used in this study allows us to perform a comprehensive analysis of the structural determinants that characterize the caspase substrate specificity. As such, we performed the prediction of three secondary-structures (helix, strand and coil), two-state solvent accessibility (exposed, buried), native disorder (disordered, ordered), as well as functional domains, using the PSIPRED (Jones, 1999), SCRATCH (Cheng *et al*., 2005), DISOPRED (Ward *et al*., 2004) and HMMER (Finn *et al*., 2008) programs, respectively. All of these programs have been widely accepted as the state-of-the-art in their respective functions.

The frequency of secondary-structure types occurring at each position from P_8 to P'_8 reveals that caspase most frequently cleave substrates that contain coils or loops, which is consistent with a recent study of large-scale proteomics-based profiling of caspase-mediated proteolytic events (Mahrus *et al*., 2008). Depending on positions P_4 through P'_4, the majority of these cleavage sites (71–80%) are observed to be located within the predicted coils, 2–9% in beta sheets and 17–19% are located within alpha-helices (Fig. 3A). However, in regards to the secondary structure (SS) motifs of cleavage sites for P_4–P'_4 sites, we found that although an all-coil motif is the most common SS motif as expected (287 counts), the second most common one is an all-helix motif (43 counts) (Supplementary Data), suggesting that selective cleavage at alpha-helical regions is not uncommon for caspases.

A large percentage of cleavages sites (65–87%, depending on P_4 through P'_4 position) are predicted to be solvent accessible (exposed), while only a small fraction of these cleavage sites (13–35%) are predicted to be solvent inaccessible (buried) (Fig. 3B). Moreover, the distribution of the predicted natively unstructured regions suggests that ∼65% of caspase cleavage sites tend to occur in natively unstructured regions (Fig. 3C), which reflects that there might exist substantial structural dynamics in the substrate-binding regions and loops or natively unstructured regions may function as distant subsites that can facilitate the caspase-substrate interactions (Garay-Malpartida *et al*., 2005). In addition, localization analysis of caspase substrate cleavage sites relative to functional domain boundaries annotated in Pfam (Finn *et al*., 2008) indicates that 35.7% of these cleavage sites are located within functional domains, 32.8% are located between functional domains, 14.0% are located before the first domain, and 17.5% are located after the last domain (Fig. 3D).

3.3 Prediction of caspase cleavage sites using the BEAA profile

In this section, we focused on predicting caspase cleavage sites from the flanking amino acid sequences of substrates. This is different from previous studies where statistical rules based on amino acid preference in the vicinity of cleavage sites (Backes *et al*., 2005; Garay-Malpartida *et al*., 2005; Lohmüller *et al*., 2003) or classification-orientated algorithms like SVM (Wee *et al*., 2006, 2007) or neural networks (Yang, 2005) are typically used to generate predictive models. Here, we formulated the prediction task of cleavage sites as a regression task for which SVR was utilized to build the models which predict the cleavage probabilities given primary sequences. First, we wanted to examine the influence of different local window sizes of single sequence inputs on the predictive performances of the SVR models. We thus used different local window sizes, both symmetric and asymmetric (Supplementary Fig. 4), to build the SVR models in order to find out which size could lead to the best performance. The accuracy of the 5-fold cross-validation is shown in Supplementary Table 2.

The predictive performance based on asymmetric windows appears to be better than the corresponding symmetric windows (Supplementary Table 2). For example, the prediction accuracy based on P_4–P'_3 (80.8%) is better than that based on P_4–P'_4 (79.6%). Another interesting finding is that the non-prime-side cleavage sites (P_1–P_8) appear to have more distant effects on the prediction performance than the prime-side sites (P'_1–P'_8) in that smaller window sizes are required to improve the predictive performance for the latter. Based on a local window of P_4–P'_2, the SVR method can achieve the best prediction accuracy of 82.2% and an MCC of 0.667. Therefore, in the following analysis, we fixed the local window at P_4–P'_2 when evaluating the effects of incorporating other informative features.

3.4 Improving predictive performance by incorporating relevant structural information of substrates

The BEAA profile used above is based on the binary encoding of single sequences only; however, we need to take into account additional informative features to further improve the predictive performance. Recently Shao *et al*. (2009) proposed a novel approach called bi-profile Bayesian feature extraction to extract the key features in the positive and negative data. This method outperformed other methods due to its advantage over binary encoding and single feature encoding schemes. On the other hand, caspase cleavage sites exhibit different features in terms of predicted secondary-structure, solvent accessibility and natively unstructured regions, which might be suitably captured by bi-profile Bayesian feature extraction approach. It is conceivable that incorporating this information might be useful for improving the prediction accuracy of caspase

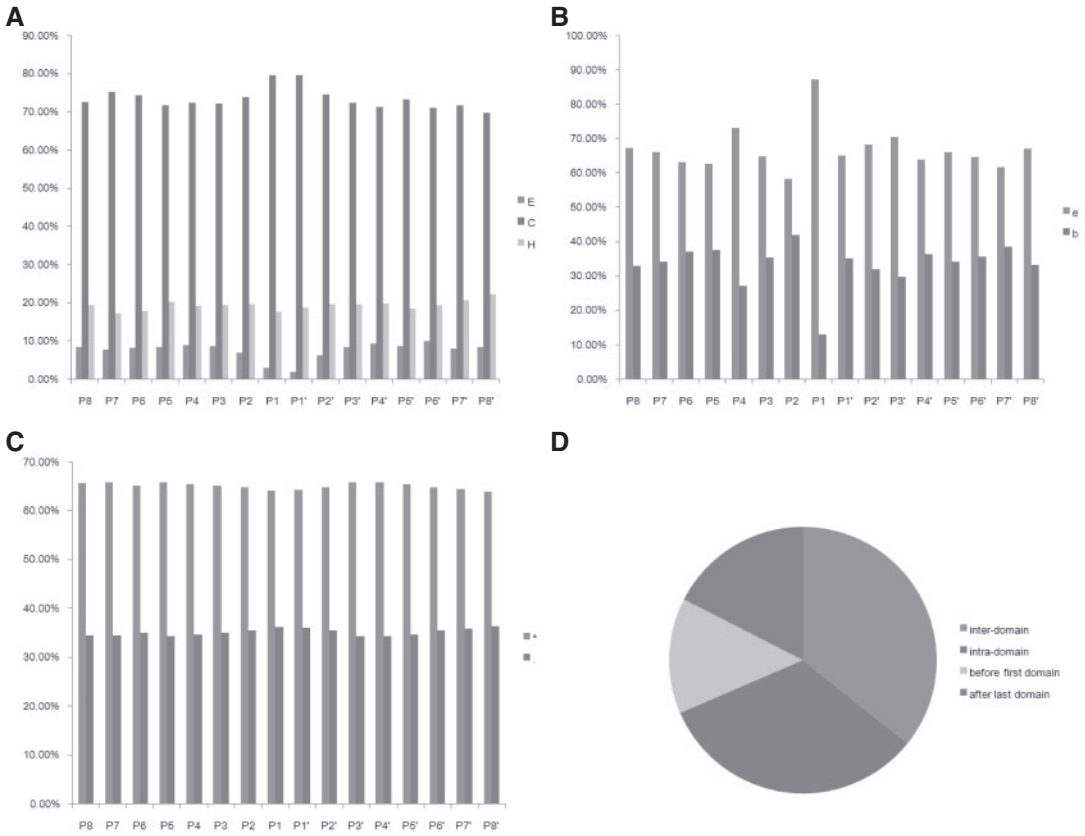

Fig. 3. Structural determinants of caspase substrate specificity based on the amino acid occurrences in P_8–P'_8 positions for cleavage sites. (**A**) Three secondary-structure assignments at each position for P_8–P'_8 positions of caspase substrate cleavage sites. H: alpha-helix, E: beta-stand, C: coil. (**B**) Two-state solvent accessibility distribution at each position for P_8–P'_8 positions of caspase substrate cleavage sites. Two states are defined: 'e' indicates 'exposed', while 'b' denotes 'buried'. (**C**) Natively unstructured region or disordered region distribution at each position for P_8–P'_8 positions of caspase substrate cleavage sites. 'Asterisk' indicates the natively unstructured/disordered region while 'dot' denotes the structured/ordered region. (**D**) Localization of caspase substrate cleavage sites relative to functional domain boundaries annotated in Pfam. Four domain boundaries are defined: inter-, intra-domain, before first domain and after last domain.

cleavage sites. From this, we extracted different types of sequence profiles based on the bi-profile Bayesian feature extraction: BPBAA, BPBSS, BPBSA and BPBDISO, as described in the 'Materials and methods' section.

We evaluated the predictive performances of different combinations of these profiles with the gradual increase in the complexity of features, i.e. the encoding schemes BEAA, BEAA_BPBAA, BEAA_BPBSA, BEAA_BPBSS_BPBSA, BEAA_BPBAA_BPBSA, BEAA_BPBAA_BPBSS_BPBSA_BPBDISO, BEAA_BPBAA_BPBSS_BPBDISO and BEAA_BPBAA_BPBSA_BPBDISO (for brevity, these encoding schemes are represented by 1, 2, …, 8, respectively, in Table 1). This step-wise procedure can reveal the contribution of individual features to the predictive performances.

When selecting the sequence encoding scheme BEAA_BPBAA which uses single binary amino acid profile coupled with bi-profile Bayesian signature, Cascleave achieved an accuracy of 84.1%, *F*-score of 74.2%, and MCC of 0.686. In contrast, adopting sequence encoding scheme BEAA_BPBSA which uses the amino acid sequence profile along with the bi-profile Bayesian solvent accessibility profile, Cascleave achieved an accuracy of 84.6%, *F*-score of 75.7%, and MCC of 0.699. Moreover, the incorporation of the bi-profile Bayesian features based on binary encoding amino acid profile and solvent accessibility (BEAA_BPBAA_BPBSA) leads to a further improvement of 1.9% accuracy. The *F*-score and MCC also increased by 4.4% and 0.031, respectively.

It is observed that the predictive performance of Cascleave gradually increases with the addition of input features, and attains the best performance when using the sequence encoding scheme BEAA_BPBAA_BPBSA_BPBDISO—achieving a prediction accuracy of 87.6%, *F*-score of 80.4% and MCC of 0.747 and was the best performing encoding scheme observed in this study. However, the addition of BPBSS did not improve the predictive performance, but rather decreases the accuracy, presumably because that there might exist redundant information between different feature sets.

To further evaluate the predictive performance, we plotted the receiver operating characteristic (ROC) curves for the assessment of these eight sequence encoding schemes in Figure 4. ROC curves indicate the performance of all the SVR models based on the same training and testing datasets. The uppermost curve with the largest area under the curve indicates the best prediction model. The ROC

Table 1. Predictive performances of Cascleave using different sequence encoding schemes

Sequence encoding schemes	Prediction accuracy (%)					
	ACC	SE	SP	*F*-score	MCC	Dimensionality
1	82.2	65.2	92.0	72.8	0.667	120
2	84.1	67.6	92.5	74.2	0.686	132
3	84.6	71.4	91.3	75.7	0.699	132
4	86.5	73.7	93.1	78.6	0.730	144
5	86.3	76.5	91.2	79.0	0.731	144
6	87.1	74.6	93.4	79.5	0.739	168
7	87.3	72.6	94.8	79.5	0.741	156
8	87.6	76.2	93.3	80.4	0.747	156

The results were obtained by 5-fold cross-validation.

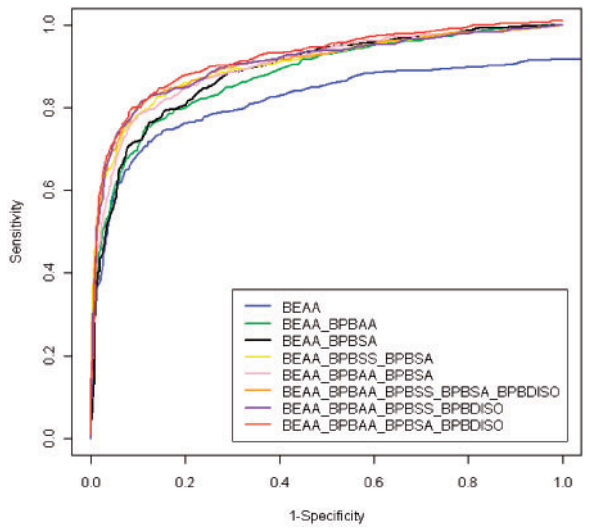

Fig. 4. The ROC curves to assess the predictive performance of the eight Cascleave models based on different sequence encoding schemes. See the main text for more details about these encoding schemes.

curve based on the scheme BEAA_BPBAA_BPBSA_BPBDISO dominates the curves representing other encoding schemes. This reinforces the notion that this is the best amongst all eight encoding schemes. Taken together, all the results obtained from Table 1, Figure 4 and Supplementary Table 5 suggest that the integration of relevant structural features in terms of predicted secondary structure, solvent accessibility and natively unstructured regions based on bi-profile Bayesian feature extraction approach can significantly enhance the predictive performance of Cascleave.

3.5 Comparison with other methods

Next, we compared the predictive performance of our Cascleave predictor with other methods. Due to the unavailability of the two previously developed webservers CasPredictor (Garay-Malpartida *et al*., 2005) and GraBCas (Backes *et al*., 2005), we compared the predictive performance between Cascleave, CASVM (Wee *et al*., 2007) and Multi-factor CASVM (Wee *et al*., 2009). Multi-factor CASVM is a two-step model: the first step is based on CASVM and the second step filters out the predicted false positives using the structural factors such as disorder and solvent exposure in the vicinity of cleavage sites (Wee *et al*., 2009). The methodological differences among these three methods are detailed in Supplementary Table 3.

Since cross-validation performance comparison is only logical when the training and testing datasets being employed are identical to each other, these two methods were first tested on the same training and testing datasets using 5-fold cross-validation. For each method, there were a total of six models indexed to a particular local window size. As previously observed by Wee *et al*. (2005), because of the large percentage of cleavage sites with the canonical 'XXXD' motif compared to the much smaller percentage of cleavage sites with the non-canonical 'XXXE' and 'XXXG' motifs, there might exist the possibility for the built models to be over-trained, meaning the built model tends to overlook the other non-canonical cleavage sites and simply predict them as being negatives. Cascleave generally results in an improvement of ∼4–5% and 2–5% with respect to CASVM and Multi-factor CASVM, respectively (Supplementary Table 4). Cascleave obtained the best performance with an accuracy of 87.4%, *F*-score of 80.3% and MCC of 0.747.

Further, we compared the predictive performance of Cascleave, CASVM and multi-factor CASVM using the LOOCV test, which is a more stringent test compared with 5-fold cross-validation (Supplementary Table 5). For the sake of a comprehensive comparison, the predictive performances of different Cascleave models based on different encoding schemes were also presented. As can be seen, Cascleave model using the encoding scheme BEAA_BPBAA_BPBSA_BPBDISO achieved the best prediction performance in the LOOCV test, with the accuracy of 89.0±2.8% and the MCC of 0.742±0.062, respectively, while multi-factor CASVM achieved the accuracy of 88.3±2.8% and the MCC of 0.715±0.072, respectively. Moreover, we tested the predictive power of the online Cascleave server to recognize novel caspase substrates and compared with the online CASVM server using an independent testing set. It was extracted from a recent study of Dix *et al*. (2008) and contains newly identified caspase substrate cleavage sites that were not previously reported. The percentage of true positives, i.e. the percentage of the cleavage sites that were correctly predicted by Cascleave is 81.2%, while CASVM was only able to identify 66.7% of the novel substrate cleavage sites, again demonstrating the predictive power of Cascleave (Supplementary Table 6).

In summary, Cascleave outperforms CASVM and multi-factor CASVM by explicitly integrating primary sequence features with the predicted solvent accessibility/secondary structures/native disorder features, which serve as an important supplement to the primary sequence. By integrating these features, Cascleave is capable of distinguishing more difficult cleavage sites that cannot be readily detected by methods based only on primary sequence information.

3.6 Case study

We further investigated the predictive performance of Cascleave by studying four different caspase substrates for which the cleavage sites have been experimentally validated. Substrate sequence scanning results with Cascleave are shown in Figure 5 and

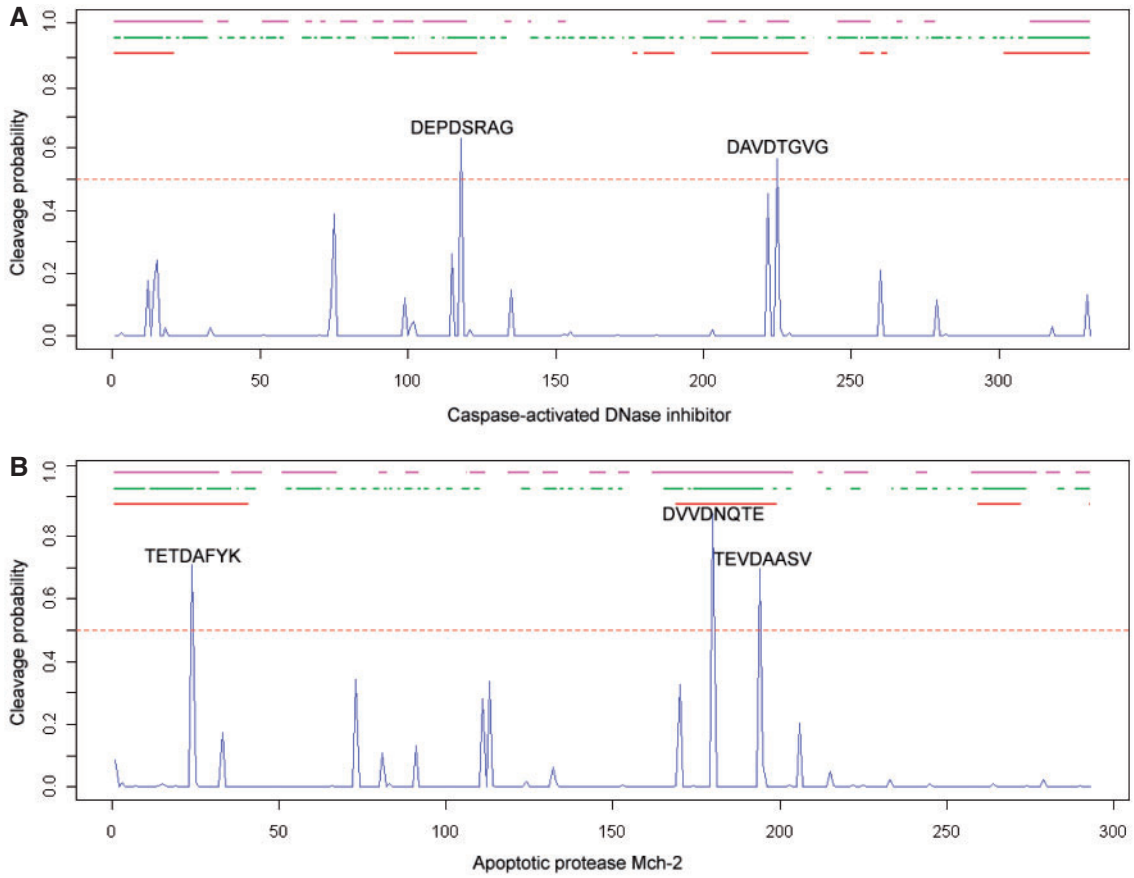

Fig. 5. The predicted cleavage probability for caspase cleavage sites using Cascleave based on the best sequence encoding scheme. (**A**) Caspase-activated DNase inhibitor (ICAD, Uniprot ID: O54786); (**B**) Apoptotic protease Mch-2 (Caspase-6, Uniprot ID: P55212). The two substrates have two and three cleavage sites, respectively. The predicted coiled, solvent exposed and natively disordered regions on the top of each panel are highlighted by magenta, green and red, respectively. A threshold value of 0.5 for making the positive cleavage site prediction is denoted by a red dashed line. The predicted cleavage sites in P_4–P'_4 positions by Cascleave are also labeled.

Supplementary Figure 5. The first example is the caspase-activated DNase inhibitor (ICAD) which inhibits the activity of Caspase-activated DNase (CAD) (Enari *et al*., 1998). Importantly, cleavage of ICAD by caspase-3 activates the CAD nuclease and deactivates its CAD-inhibitory effect. Site-directed mutagenesis indicated that ICAD carries two specific cleavage sites by caspase-3: DEPD|SRAG at positions 117–118 and DAVD|TGVG at positions 224–225 ('|' denotes the cleavage site). Cleavage of ICAD can liberate the active CAD nuclease that can otherwise be inhibited by ICAD and mediate the apoptotic DNA degradation (Sakahira *et al*., 1998).

The second example is the apoptotic protease Mch-2 (caspase-6), which is further cleaved by caspase-3, -8 or -10 to produce the two active subunits (Srinivasula *et al*., 1996), resulting in the activation cascade of caspases responsible for apoptosis execution. Mch-2 has three cleavage sites of capase-3: TETD|AFYK at positions 23–24, DVVD|NQTE at position 179–180 and TEVD|AASV at positions 193–194 (Srinivasula *et al*., 1996). All the experimentally identified cleavage sites of both ICAD and Mch-2 were successfully predicted with the Cascleave predictor based on the best-performing sequence encoding scheme (Fig. 5). These cleavage sites were among the top ranking results according to the predicted probabilities and could be correctly identified as putative cleavage sites using a cutoff threshold of 0.5.

The third and fourth examples are the heterogeneous nuclear ribonucleoprotein inhibitor (hnRNP, Uniprot ID: O43390) (Brockstedt *et al*., 1998) and the Ras GTPase-activating protein (RasGAP, Uniprot ID: P20936) (Yang and Widmann, 2001), respectively. They represent difficult caspase substrates for which the cleavage sites cannot be readily deduced from the use of sensitive machine learning models like Cascleave. hnRNP has four caspase cleavage sites, however, Cascleave only successfully predicted two of them: DYYD|DYYG and DYHD|YRGG, failing to identify the other two cleavage sites: RAID|ALRE and KESD|LSHV (Supplementary Fig. 5). RasGAP is a special substrate as it is cleaved through sequential caspase cleavage during apoptosis (Yang and Widmann, 2001). Assigning a higher cutoff threshold of 0.6, Cascleave was only able to predict one of the two cleavage sites DEGD|SLDG, while missing the primary cleavage site DTVD|GKEI. Nevertheless, if a lower cutoff of 0.5 is used, DTVD|GKEI site will be included in the predicted sites, with the ranking of third place, among the five predicted sites. All the above results suggest that silico computational sequence scanning using

Cascleave might be helpful for identifying the putative caspase substrate cleavage sites *in vivo*.

4 DISCUSSION

Caspases have central roles in apoptotic cell death processes by catalysing a multitude of proteolytic events, such as activating pro-survival or anti-apoptotic proteins and activating anti-survival or pro-apoptotic proteins (Mahrus *et al*., 2008). The key to understanding the physiological role of caspases is to identify their natural substrates. Caspases have the ability to cleave multiple proteins in different physiological compartments and produce polypeptide products with potentially new functional activities (Dix *et al*., 2008; Lüthi and Martin, 2007; Mahrus *et al*., 2008), with cleavage preference influenced by various factors such as substrate sequence, substrate conformation and accessibility. Knowledge about the substrate specificity of caspases can dramatically improve our ability to predict target protein substrates; however, this information cannot at present be readily derived from experimental approaches. Solving this problem is fundamental for both understanding caspase biology and the development of therapeutics that target specific caspase-mediated apoptotic pathways.

The recent proliferation of large-scale *in vitro* peptide cleavage libraries and proteome-wide global profiling platforms offers promising prospects of identifying specific caspase substrates in combination with computer-based screening of genome sequences (Timmer *et al*., 2009). The advantage of machine learning techniques such as SVR make them particularly appealing in solving the difficult problem of caspase substrate specificity prediction, as they can be effectively applied to better describe the complex non-linear relationships underlying the caspase-substrate interactivity by using the kernel functions to build the predictive models (Shao *et al*., 2009). To address this problem, we developed a novel bioinformatic approach based on SVR to make testable predictions on the substrate specificity of caspases. Considering that caspase cleavage sites show preferences for predicted secondary structure, solvent accessibility and natively unstructured regions, we used a bi-profile Bayesian feature extraction approach to derive these profiles and train the SVR models. The efficiency of the resulting Cascleave predictors for predicting caspase cleavage sites has been demonstrated by comparing to one of the most accurate existing algorithms CASVM (Wee *et al*., 2006, 2007) and multi-factor CASVM (Wee *et al*., 2009). The predictive performance of Cascleave was further showcased by predicting four caspase substrates for which the cleavage sites have been experimentally validated. The results indicate that sequence scanning using Cascleave should be very useful for identifying the putative caspase cleavage sites.

In this study, our goal was to predict all the potential cleavage sites, irrespective of the spatiotemporal environments or conformation changes that a substrate may be subject to. The features used as input to the Cascleave predictor are derived from primary sequences and the input data does not contain specific information about the order of sequential cleavage events in proteolytic cascades. In these exceptional cases, there is a chance that the predictor will fail as a purely statistical model. The predicted solvent exposure and native disorder features are a supplement to the primary sequence, and in cases where the primary sequence alone contains a strong consensus indicating a cleavage site, we hope these amino acid sequence encoding features are overriding. For example, in RasGAP (Uniprot ID: P20936) there is a correctly predicted cleavage site (DEGD|SLDG) by Cascleave which shows low probability of being solvent exposure and natively disordered (Supplementary Fig. 5). In the specific cases where a sequential proteolytic cascade occurs, features based on primary sequence alone are unlikely to be sufficient. Training based on features from the 3D structure of the substrate protein may help in this regards, although this is not amenable to a general predictor since the 3D structures of many substrates, in their intact or cleaved-form mid-cascades, are generally not known.

There are a number of additional measures with potential to further increase the accuracy of prediction in the future. The first approach is to use either the accurate 3D structure of the substrate or the structure of a protease in complex with a substrate, instead of the predicted structural information to build the predictive models. This will allow us to ascertain whether the high-resolution crystal structure of a protease/substrate complex can be used to derive specificity information. The second is to incorporate other informative and complementary features, such as sequence-order-dependent context that can better describe the sequential neighborhood surrounding the substrate cleavage sites (Mahrus *et al*., 2008; Nicholson, 1999). The third is to investigate how to effectively represent the negatives (non-cleavage sites) that are true negatives and are non-cleavable under any cellular or physiological conditions, and how to better discriminate these absolute negatives from those that are cleavable under a given physiological condition. Thus the precise detection of the true negatives is more likely to contribute to an improvement of the prediction accuracy, based on which characteristic feature sets regarding positives and negative samples can be more accurately represented and established. Further improvement can be also achieved by using refined training and testing datasets with high-quality coverage of the cleavage sites, identified by high-throughput proteome-wide techniques.

5 CONCLUSION

In summary, we have proposed a novel approach to predict cleavage sites from the flanking amino acid sequences of caspase substrates. We analyzed the structural determinants of caspase substrate specificity based on a well-curated database from multiple resources. We built the Cascleave models using the bi-profile Bayesian approach that takes into account the characteristic sequential and structural profiles in the vicinity of cleavage and non-cleavage sites. This provides a quantitative evaluation of caspase substrate specificity. The results showed that our approach is more accurate than the CASVM and multi-factor CASVM methods, demonstrating its usefulness for describing the complex sequence-structure relationships and predicting caspase cleavage sites. It is anticipated that Cascleave will be a powerful tool for predicting novel substrate cleavage sites of caspases and shedding new insights on the unknown caspase-substrate interactivity.

Funding: J.S. would like to thank the National Health and Medical Research Council of Australia (NHMRC) and the Japan Society for the Promotion of Science (JSPS) for financially supporting this research via the NHMRC Peter Doherty and JSPS Postdoctoral Fellowships. H.S. was supported by the National

Natural Science Foundation of China (60704047), Science and Technology Commission of Shanghai Municipality (08ZR1410600, 08JC1410600), and sponsored by Shanghai Pujiang Program and Innovation Program of Shanghai Municipal Education Commission (10ZZ17). G.I.W. is supported by the Australian Research Council (DP0772238). J.C.W. is a NHMRC Principal Research Fellow and an ARC Federation Fellow.

Conflict of Interest: none declared.

REFERENCES

Ahmed,F. *et al*. (2009) Prediction of guide strand of microRNAs from its sequence and secondary structure. *BMC Bioinformatics*, **10**, 105.

Backes,C. *et al*. (2005) GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Res.*, **33**, W208–W213.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Brockstedt,E. *et al*. (1998) Identification of apoptosis-associated proteins in a human Burkitt lymphoma cell line. Cleavage of heterogeneous nuclear ribonucleoprotein A1 by caspase 3. *J. Biol. Chem.*, **273**, 28057–28064.

Brown,J.B. and Akutsu,T. (2009) Identification of novel DNA repair proteins via primary sequence, secondary structure, and homology. *BMC Bioinformatics*, **10**, 25.

Cheng,J. *et al*. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res*., **33**, W72–W76.

Denecker,G. *et al*. (2008) Caspase-14 reveals its secrets. *J. Cell Biol.*, **180**, 451–458.

Dix,M.M. *et al*. (2008) Global mapping of the topography and magnitude of proteolytic events in apoptosis. *Cell*, **134**, 679–691.

Enari,M. *et al*. (1998) A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD. *Nature*, **391**, 43–50.

Enoksson,M. and Salvesen,G.S. (2008) Proteolytic needles in the cellular haystack. *Nat. Chem. Biol.*, **4**, 651–652.

Enoksson,M. *et al*. (2007) Identification of proteolytic cleavage sites by quantitative proteomics. *J. Proteome Res.*, **6**, 2850–2858.

Finn,R.D. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res*., **36**, D281–D288.

Fischer,U. *et al*. (2003) Many cuts to ruin: a comprehensive update of caspase substrates. *Cell Death Differ*., **10**, 76–100.

Garay-Malpartida,H.M. *et al*. (2005) CaSPredictor: a new computer-based tool for caspase substrate prediction. *Bioinformatics*, **21**, i169–i176.

Gasteiger,E. *et al*. (2005) Protein identification and analysis tools on the ExPASy server. In Walker,J.M. (ed.). *The Proteomics Protocols Handbook*. Humana Press, Totowa, New Jersey, pp. 571–607.

Joachims,T. (1999) Making large-Scale SVM learning practical. In Schölkopf,B. *et al*. (eds). *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*., **292**, 195–202.

Ju,W. *et al*. (2007) Proteome-wide identification of family member-specific natural substrate repertoire of caspases. *Proc. Natl Acad. Sci. USA*, **104**, 14294–14299.

Kalita,M.K. *et al*. (2008) CyclinPred: a SVM-based method for predicting cyclin protein sequences. *PLoS ONE*, **3**, e2605.

Kumar,M. and Raghava,G.P. (2009) Prediction of nuclear proteins using SVM and HMM models. *BMC Bioinformatics*, **10**, 22.

Liu,W. *et al*. (2006) Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics*, **7**, 182.

Lohmüller,T. *et al*. (2003) Toward computer-based cleavage site prediction of cysteine endopeptidases. *Biol. Chem.*, **384**, 899–909.

Lüthi,A.U. and Martin,S.J. (2007) The CASBAH: a searchable database of caspase substrates. *Cell Death Differ.*, **14**, 641–650.

Mahrus,S. *et al*. (2008) Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell*, **134**, 866–876.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Nicholson,D.W. (1999) Caspase structure, proteolytic substrates, and function during apoptotic cell death. *Cell Death Differ.*, **6**, 1028–1042.

Pop,C. and Salvesen,G.S. (2009) Human caspases: activation, specificity and regulation. *J. Biol. Chem.*, **284**, 21777–21781.

Raghava,G.P. and Han,J.H. (2005) Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC Bioinformatics*, **6**, 59.

Rawlings,N.D. *et al*. (2008) MEROPS: the peptidase database. *Nucleic Acids Res.*, **36**, D320–D325.

Sakahira,H. *et al*. (1998) Cleavage of CAD inhibitor in CAD activation and DNA degradation during apoptosis. *Nature*, **391**, 96–99.

Schilling,O. and Overall,C.M. (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nature Biotechnol.*, **26**, 685–694.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Shao,J. *et al*. (2009) Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS ONE*, **4**, e4920.

Shen,J. *et al*. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.

Singh,G.P. (2006) Intrinsic unstructuredness and abundance of PEST motifs in eukaryotic proteomes. *Proteins*, **62**, 309–315.

Song,J. and Burrage,K. (2006) Predicting residue-wise contact orders in proteins by support vector regression. *BMC Bioinformatics*, **7**, 425.

Song,J. *et al*. (2006) Prediction of *cis*/*trans* isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC Bioinformatics*, **7**, 124.

Song,J. *et al*. (2007) Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics*, **23**, 3147–3154.

Song,J. *et al*. (2008) HSEpred: predict half-sphere exposure from protein sequences. *Bioinformatics*, **24**, 1489–1497.

Song,J. *et al*. (2009) Prodepth: predict residue depth by support vector regression approach from protein sequences only. *PLoS ONE*, **4**, e7072.

Srinivasula,S.M. *et al*. (1996) The Ced-3/interleukin 1beta converting enzyme-like homolog Mch6 and the lamin-cleaving enzyme Mch2alpha are substrates for the apoptotic mediator CPP32. *J. Biol. Chem*., **271**, 27099–27106.

Talanian,R.V. *et al*. (1997) Substrate specificities of caspase family proteases. *J. Biol. Chem.*, **272**, 9677–9682.

Tamura,T. and Akutsu,T. (2007) Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. *BMC Bioinformatics*, **8**, 466.

Timmer,J.C. and Salvesen,G.S. (2007) Caspase substrates. *Cell Death Differ*., **14**, 66–72.

Timmer,J.C. *et al*. (2009) Structural and kinetic determinants of protease substrates. *Nat. Struct. Mol. Biol.*, **16**, 1101–1108.

Vapnik,V. (2000) *The Nature of Statistical Learning Theory*. Springer, New York.

Wan,J. *et al*. (2006) SVRMHC prediction server for MHC-binding peptides. *BMC Bioinformatics*, **7**, 463.

Ward,J.J. *et al*. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Wee,L.J. *et al*. (2006) SVM-based prediction of caspase substrate cleavage sites. *BMC Bioinformatics*, **7**(Suppl. 5), S14–S15.

Wee,L.J. *et al*. (2007) CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. *Bioinformatics*, **23**, 3241–3243.

Wee,L.J. *et al*. (2009) A multi-factor model for caspase degradome prediction. *BMC Genomics*, **10**(Suppl. 3), S6.

Yang,J.Y. and Widmann,C. (2001) Antiapoptotic signaling generated by Caspase-induced cleavage of RasGAP. *Mol. Cell. Biol*., **21**, 5346–5358.

Yang,Z.R. (2005) Prediction of caspase cleavage sites using Bayesian bio-basis function neural networks. *Bioinformatics*, **21**, 1831–1837.

Yuan,Z. (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, **6**, 248.